УДК 004.63

Л. А. Савицька, Т. І. Коробейнікова, П. В. Чирва

МЕТОД ТА КРОСПЛАТФОРМЕНИЙ ЗАСІБ АРХІВАЦІЇ ОДНОТИПНИХ ФАЙЛІВ

Вінницький національний технічний університет, Вінниця

Анотація. Дана робота присвячена розробці методу та кросплатформеній реалізації засобу архівації однотипних файлів. Цей програмний засіб дозволить ефективно виконувати архівацію великої кількості однотипних файлів.
Високий рівень вирішення поставленої задачі досягнуто за рахунок використання сучасної мови програмування Java.
В даній роботі виконано дослідження і аналіз сучасних методів та засобів архівації даних в галузі інформаційних технологій, аналіз підходів, методів та моделей архівації даних. Розглянуті методи стиснення та архівації даних та їх порівняльна характеристика. Виконано огляд сучасних засобів архівації даних, а саме, процеси архівації за кросплатформеною технологією Java. Дослідження, виконані під час дослідження, ґрунтуються на теоретико-множинних підходах і принципах кросплатформеного підходу для виконання процесів архівації за кросплатформеною технологією Java; структурному проектуванні програмного забезпечення – для реалізації кросплатформеного програмного забезпечення архівації однотипних файлів; методах об'єктно-орієнтованого програмування – для реалізації алгоритмів і процесів методу архівації однотипних файлів та розробки відповідного програмного забезпечення.
Зокрема, у роботі розроблено метод архівації однотипних файлів, вдосконалено процес формування основного словника, вдосконалено процес архівації файлів.
Ключові слова: метод та кросплатформена реалізація засобу архівації однотипних файлів, методи стиснення та архівації даних, кросплатформена технологія Java, ключовий файл, початковий словник, однотипні файли, стиснення з втратами, стиснення без втрат.
Abstract. This work is devoted to the development of a method and cross-platform implementation of a file archiving tool of the same type. This software tool will allow you to efficiently archive a large number of files of the same type.
The high level of the solution to this task was achieved through the use of modern Java programming language.
In this thesis research and analysis of modern methods and means of archiving of data in the field of information technologies, analysis of approaches, methods and models of data archiving have been performed. The methods of compression and archiving of data and their comparative characteristics are considered. An overview of modern data archiving tools is performed, namely, processes of archiving on a cross-platform Java technology. The research performed during the research is based on multiple-theoretical and cross-platform approaches for performing cross-platform archiving of Java technology; structural software design - for implementation of cross-platform archiving of the same files; object-oriented programming methods - to implement algorithms and processes of the method of archiving of the same files and development of the corresponding software.
In particular, the method of archiving of the same type files has been developed, the process of formation of the main vocabulary has been improved, the file archiving process has been improved.
Keywords: method and cross-platform implementation of the same file archiving tool, data compression and archiving methods, cross-platform Java technology, key file, source dictionary, single file, lossy compression, lossless compression.
Аннотация. Данная работа посвящена разработке метода и кроссплатформенных реализации средства архивации однотипных файлов. Этот про- программно средство позволит эффективно выполнять архивацию большого количества однотипных файлов.
Высокий уровень решения поставленной задачи достигнуто за счет использования современного языка программирования Java.
В данной работе выполнено исследование и анализ современных методов и средств архивации данных в области информационных технологий, анализ подходов, методов и моделей архивации данных. Рассмотрены методы сжатия и архивации данных и их сравнительная характеристика. Выполнен обзор современных средств архивации данных, а именно, процессы архивации по кросплатформеною технологии Java. Исследования, выполненные в ходе исследования, основанные на теоретико-множественных подходах и принципах кроссплатформенных подхода для выполнения процессов архивации по кросплатформеною технологии Java; структурном проектировании программного обеспечения - для реализации кроссплатформенных программного обеспечения архивации однотипных файлов; методах объектно-ориентированного программирования - для реализации алгоритмов и процессов метода архивации однотипных файлов и разработки соответствующего программного обеспечения.
В частности, в работе разработан метод архивации однотипных файлов, усовершенствован процесс формирования основного словаря, усовершенствован процесс архивации файлов.
Ключевые слова: метод и кроссплатформенных реализация средства архивации однотипных файлов, методы сжатия и архивации данных, кросплатформена технология Java, ключевой файл, начальный словарь, однотипные файлы, сжатие с потерями, сжатие без потерь.
DOI: https://doi.org/10.31649/1999-9941-2020-47-1-14-21.

Вступ

Задача компактного зберігання, перетворення та передавання інформаційних даних завжди була актуальною в галузі інформаційних технологій.

Інформаційні ресурси нині є продуктом інтелектуальної діяльності дійсно найбільш кваліфікованої й творчо активної частини молоді та працездатного населення світу. В останній чверті XX століття набуті інформаційні ресурси досягли (і продовжують досягати) настільки рекордних обсягів, що цілком повсякденним стали поняття «інформаційного вибуху», «інформаційної революції». В якості доказу є об'єктивне збільшення інформаційного потоку з початку цього сторіччя більш ніж в 30 разів! [1].

Актуальність

Отже, **актуальною** є наукова задача розробки та застосування принципово нових методів і засобів сприйняття, передачі, обробки, зберігання і розповсюдження інформаційних даних, таких, що здатних оперувати великими масивами інформації, причому, у реальному часі.

З метою забезпечити надійне збереження інформації створюють резервні копії даних. Задача збереження резервних копій у компактному вигляді є основою для процесів архівації та стиснення даних. В загальному випадку, основний зміст архівації полягає у створенні таких резервних копій, які потребували би значно меншого обсягу на інформаційних ресурсах, ніж та сама інформація у вихідному стані. Таким чином, в контексті під архівацією слід розуміти процес перекодування деякої сукупності файлів з метою зменшення загального об'єму пам'яті, який вони займають. Часто архівацією ще називають процес стиснення даних [3].

Нині відомо досить багато різних підходів до процесу архівації. Усі підходи мають в своїй основі різні підходи та різні методи, проте подібні вони в одному – це те, що вони сповідують принцип заміни рівномірного двійкового коду на нерівномірний. З метою архівації файлів та папок використовують спеціальні програмні засоби, які називають архіваторами. Стиснуті файли поміщають у файли, який називають архівами [2].

Перші прототипи архіваторів з'явилися у 80-х роках минулого сторіччя. Основними можливостями сучасних архіваторів є такі:

- занесення груп файлів та (або) підкаталогів в архів;
- можливість поновлення архіву;
- перегляд файлів з меж архіву;
- вилучення окремих файлів з архіву;
- захист файлів від несанкціонованого доступу (НСД);
- перевірка архіву на цілісність;
- створення багатотомних архівів;
- можливість створення архівів, що автоматично відкриваються.

Можливості сучасних програм-архіваторів дозволяють зекономити від 20 до 90 відсотків дискового простору. Файлом, що знаходиться в архіві, можна скористатися після того, як він буде відновлений у початковому вигляді, тобто розархівований (розпакований). Розархівування виконують або ті ж самі програми-архіватори в зворотному напрямку, або окремі програми, які називають розраховаторами, серед яких найбільш відомими є: ZIP, JAR, RAR. Під час вибору конкретного засобу для архівування (розархівування) користувачі керуються багатьма критеріями, як то швидкістю роботи, коефіцієнти стискування даних, інтерфейс, сумісність тощо. Важливим є те, що для одного типу файлів кращим може бути один архіватор, а для іншого – інший [5].

Мета

Метою дослідження є збільшення середнього значення процесу архівації для великої кількості однотипних файлів.

Для досягнення поставленої мети необхідно виконати такі завдання:

- провести аналіз сучасних методів та засобів архівації даних в галузі інформаційних технологій, виконати їх порівняльну характеристику та сформувати вимоги та обрати й обґрунтувати вибір методу, що задовольняв би меті даного дослідження;
- розглянути існуючі способи архівації за кросплатформеною технологією Java;
- запропонувати метод архівації однотипних файлів згідно мети магістерської кваліфікаційної роботи, розробити ключові процеси роботи методу архівації однотипних файлів та виконати програмну реалізацію запропонованого методу архівації однотипних файлів;
- провести тестування програмного продукту та виконати аналіз отриманих результатів.

Сучасні методи архівації даних в галузі інформаційних технологій

Характерною особливістю усіх наявних типів інформаційних даних є їх надлишковість. Для людини, як для суб'єкта, надлишковість цих даних часто пов'язана із якістю отриманої даних, оскільки така надлишковість покращує нам зрозумілість та сприйняття цієї даних. Проте, коли ми говоримо про зберігання та передавання даних засобами обчислювальної техніки, то наявність надлишковості відіграє дуже негативну роль, оскільки вона призводить до зростання вартості зберігання та передавання даних [4].

Особливо актуальною є ця задача у випадках необхідності оброблення величезних обсягів даних при незначних чи недостатніх об'ємах носіїв даних. У зв'язку із цим постійно виникає задача відійти від надлишковості або процесу стиснення та архівації даних.

Базовий принцип, що є основою для процесу стиснення та архівації даних, полягає в економічному описанні повідомлення, згідно якого можливе відновлення його початкового значення із похибкою, яка контролюється [4].

Аналіз підходів, методів та моделей архівації даних

Методи стиснення та архівації даних можна розділити на два типи:

1) без спотворення (loseless) – методи стиснення та архівації гарантують, що декодовані дані будуть в збігатися із вихідними;
2) із втратами (lossy) – методи процесу стиснення та архівації можуть спотворювати вихідні дані, за рахунок видалення несуттєвих частин даних, після чого повне відновлення неможливе .

Методи стиснення та архівації даних без втрат засновані на усуненні надлишковості подання даних. Ефективність кодування досягається за рахунок подання малоймовірних подій більш довгими словами, ніж подій із вищою ймовірністю настання.

Якщо ймовірність настання події є деяке значення P, то, відповідно теоремі Шеннона, цю подію варто кодувати словом завдовжки -log2P біт. Методи стиснення та архівації даних явно або неявно мають в основі саме це [4].

У результаті процесів ефективного кодування «1» вихідних даних ставиться у відповідність КС (КС). КС складається із послідовності двійкових цифр. Сукупність кодових слів утворює сам код. У випадку, коли довжини всіх кодових слів сталі, то застосовується код має фіксовану (постійну) довжину, інакше – змінну. Якщо вихідні дані можуть бути відновлені по масиву кодових слів, то кодування не призводить до втрат даних.

Ефективність процесів стиснення та архівації визначається ступінню стиснення. Ступінь самого стиснення становить значення, яке дорівнює відношенню обсягу вихідних даних до об'єму відповідних їм стиснутих даних і вимірюється в кількості разів.

Всі методи стиснення та архівації прийнято розділяти на 2 класи:

1) статистичного кодування
2) словникового стиснення та архівації [2].

У схемах процесів стиснення та архівації часто застосовуються допоміжні перетворення, що забезпечують та сприяють виконанню ефективного кодування.

Метод та кросплатформений засіб архівації однотипних файлів

Розроблений метод призначений для процесів оптимізованої архівації великих кількостей маленьких однотипних файлів. В основі роботи запропонованого методу архівації однотипних файлів є ідея заміни повторного входження цілого блока даних посиланням на попередню позицію його входження.

Загальна схема складових та процесів методу архівації однотипних файлів зображена на рис. 1.

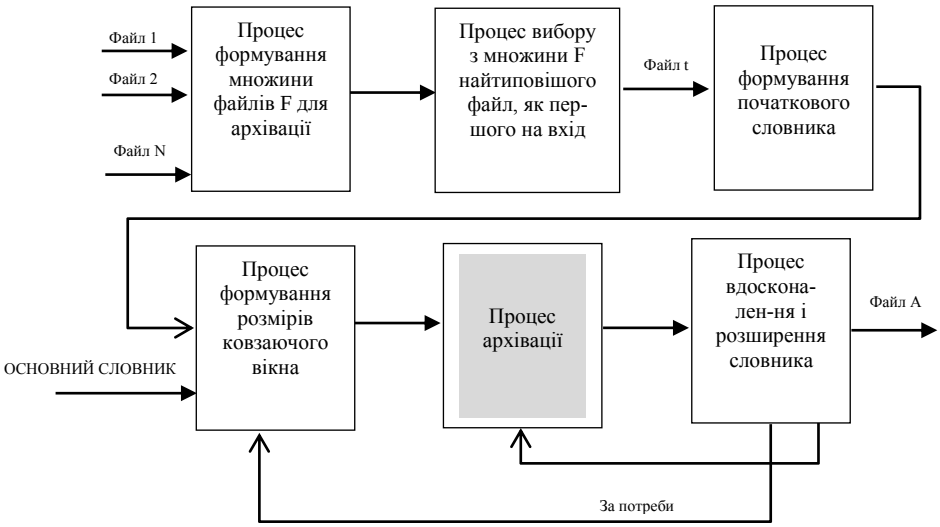

Рисунок 1 – Загальна схема методу архівації однотипних файлів

Метод архівації однотипних файлів передбачає забезпечення та виконання таких процесів:

1) Прийом на вхід множини файлів;
2) Процес формування множини файлів F для архівації;
3) Процес вибору з множини F найтиповішого файл, як першого на вхід;

4) Виокремлення такого ключового файлу;
5) Процес формування початкового словника;
6) Процес формування розмірів ковзаючого вікна. «Ковзаюче вікно» в даному випадку є динамічною структурою даних, що організована таким чином, аби містити в собі введену раніше інформацію та надавати до неї доступ;
7) Процес архівації. Передбачає звертання до елементів «ковзаючого вікна» і замість значень послідовності, що архівується, вставлення посилань на ці значення своєрідному «словнику»;
8) За потреби – перехід до процесу формування розмірів ковзаючого вікна.
9) Процес вдосконалення і розширення словника. Перехід на процес архівації знову;
10) Архівний файл А.

3.1 Розробка процесу формування початкового словника

Процес формування словника у стандартному використанні алгоритму LZ77 це формування комбінації, що повторюється і кодуються парою:

– довжина збігу (match length)
– зсув (offset) або дистанція (distance).

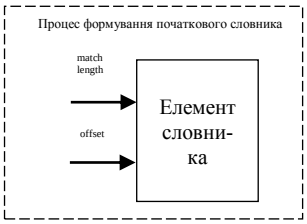

Рисунок 2 – Процес формування словника

Під час процесу формування початкового словника кожна така комбінація «довжина збігу + зсув» трактується як команда копіювання символів із певної позиції «ковзаючого вікна», або дослівно звучить так: «Повернутися назад в словнику на значення «зсуву» символів і скопіювати значення «довжини збігу» символів, починаючи із поточної позиції».

3.2 Розробка процесу архівації

Особливість розглянутого в даному методі архівації однотипних файлів є сам процесу архівації, який полягає в тому, що використання комбінації кодової пари «довжина-зміщення» є не тільки прийнятним, але й ефективним у тих випадках, коли значення довжини збігу *match length* перевищує значення зсуву *offset*.

В кінцевому підсумку, ступінь ефективності процесу та архівації в даному методі архівації однотипних файлів залежить від того, наскільки багато в файлі повторюваних комбінацій, і наскільки вони великі.

Це приводить нас до такої запропонованої ідеї: з метою збільшити кількість таких блоків, необхідно «об’єднати» всі файли, які планується архівувати, в один великий файл, і вже після цього його, цей великий файл, стискати. Звісно, такий підхід буде дуже ефективним для однотипних файлів (і виправдовує мету даної роботи), і вкрай неефективним для різних типів файлів, іншими словами, із різним профілем даних. Тому застосування запропонованого методу архівації однотипних файлів може давати високі результати роботи процесу архівації.

Оскільки запропонований метод тісно пов’язаний із змістом даних, що архівуються, можна говорити про ентропію.

В загальному випадку, дані, що отримуються приймачем несуть корисну інформацію, якщо має місце невизначеність відносно стану джерела даних. Величина, що визначає невизначеність окремого і-го повідомлення називають частковою ентропією (1).

$$M(x_i) = \frac{log_2 y}{p(x_i)} \quad (1)$$

І тоді кількість інформації і невизначеність для цієї множини випадкових повідомлень можна отримати шляхом усереднення по всіх елементам даних (2)-(3).

$$\mathrm{I}(x) = -\sum p(x_i) * \frac{log_a 1}{p(x_i)}; \tag{2}$$

$$H(x_i) = -\sum P(x_i) log_a p(x_i) \tag{3}$$

де $H(x_i)$ – ентропія.

Не дивлячись на те, що є співпадіння залежностей, все ж ентропія і кількість інформації є принципово різними. Сама ентропія, що виражає «середню невизначеність джерела» даних є характеристикою джерела даинх і, якщо нам доступна статистика повідомлень, яка може бути визначена заздалегідь. $I(x)$ є попередньою характеристикою і визначає кількість даних, що отримуються із вхідного повідомлення. $H(x)$ – це, так звана, міра нестачі інформації про стан окремої частини якоїсь системи чи системи. Пропорційно надходженню даних про стан системи, ентропія останньої стає меншою. Співпадіння виразів (2) і (3) говорить про те, що кількість отриманих даних в числововму еквіваленті дорівнює ентропії, що існує залежно від джерела даних на розглянутій частині каналу зв'язку тією кількості інформації з ентропією, що чітко проявляється діалектичний закон.

Середнє значення ентропії даних при однаковій кількості елементів може відрізнятися залежно від статистики і характеристик самих даних. Під час наявності зв'язків залежності між елементами даних, ентропія стає меншою. У випадках, коли зв'язки залежності охоплюють 2 або 3 елементи, тоді формула для обчислення ентропії стане (4) для 2 елементів і (5) для 3 елементів.

$$H(x) = -\sum\sum p(x_i, x_j) log_a p(x_i, x_j), \tag{4}$$

$$H(x) = -\sum\sum\sum p(x_i, x_j, x_k) log_a p(x_i, x_j, x_k) \tag{5}$$

Ті початкові дані є кращими, у яких показники ентропії є оптимальними. Виміром наскільки дані за своєю ентропією відрізняється оптимальних даних, покаже коефіцієнт архівації (6):

$$\mu = \frac{H(x)}{H(x)_{max}} \tag{6}$$

де $H(x)$ – реальне; $H(x)_{max}$ – ентропія відповідному йому оптимального повідомлення.

Якщо дані, що не є оптимальними і оптимальні деякі дані характеризуються однаковим значенням ентропієї, тоді справедлива така рівність (7)

$$n * H(x) = n' * H(x)\max, \tag{7}$$

де n – число елементів не оптимального повідомлення даних; n' – число елементів відносно оптимального повідомлення даних.

Оскільки ентропія для оптимальних даних є максимальною, тоді число елементів неоптимальних даних n завжди буде більшою від кількості елементів відповідних оптимальних даних n'. І тоді коефіцієнт ахівування можна виразити через кількість елементів повідомлення (8)

$$\mu = \frac{n'}{n}. \tag{8}$$

Так, реальні дані тоді при однаковій ступені інформативності володіють визначеною надлишковістю в своїх елементах у порівнянні з оптимальними даними. Дійсним коефіцієнтом архівування вважатимемо (9)

$$K_A = \frac{N_{in}}{N_{out}}. \tag{9}$$

де N_{in} – кількість двійкових розрядів на вході методу архівації. N_{out} – кількість двійкових розрядів на виході методу архівації.

В рамках даної роботи це було експериментально досліджено автором, і виявлено, що текстові файли були заархівовані гірше за файли реляційної бази даних. Саме цей факт пояснює тим, що в текстових (чи інших складних форматів) файлах, наявне досить незначне число повторюваних блоків в цих файлах.

В таблиці 1 наведено експериментальні дані архівації однотипних файлів по одному та однотипних файлів у кількості 10 шт. та 100 шт.

Середнє значення процесу архівації для різних типів файлів Arh_m розраховане для процесу архівації для різних типів файлів, кількість – 1 файл, процесу архівації для однотипних файлів, кількість – 10 файлів та процесу архівації для однотипних файлів, кількість – 100 файлів.

Таблиця 1 – Середнє значення процесу архівації

Тип файлів	Arh_m – 1 файл, %	Arh_m – 10 файлів, %	Arh_m – 100 файлів, %
Документ (*.docx)	62,875	40,33	47,4
Документ (*.doc)	37,7825	32, 567	35,7
Текстовий (*.txt)	55,305	47, 376	56,78
Бази даних (*.accdb)	18,89	83,02	85,189
Графічні файли (*.jpg)	27,31	28,5	30,01
Звукові файли (*.mp3)	76,4	77,08	78,5
HTML (*.html)	14,74	24,129	35,67
Усереднене значення показника архівування	41,9003571	50,612	52,75

На рис. 3 наведено графічне представлення результатів дослідження процесів архівації і виявлено, що розроблений метод архівації однотипних файлів гарно працює в умовах наявності не менше 10 файлів на вході методу, а при наявності більшої кількості (100 і більше) однотипних файлів значення Arh_m росте.

а)

Рисунок 3 – Діаграма процесу архівації

Запропонований метод архівації однотипних файлів дає приріст середнього значення процесу архівації на 10,85%. А зі збільшенням кількості однотипних файлів для процесу архівації цей показник може все більше зростати.

Висновки

Підсумком виконання даного дослідження роботи стала розробка методу та кросплатформеного засобу архівації однотипних файлів. Високий рівень виконання поставленої прикладної задачі вирішено засобами мови програмування Java.

Розроблений метод призначений для процесів оптимізованої архівації великих кількостей маленьких однотипних файлів. В основі запропонованого роботи методу архівації однотипних файлів є ідея заміни повторного входження цілого блока даних посиланням на попередню позицію його входження.

Зокрема, у роботі отримані такі наукові результати:

- вперше запропоновано метод архівації однотипних файлів, який дозволяє збільшити середнє значення процесу архівації однотипних файлів на 10, 85%;
- вдосконалено процес формування основного словника за рахунок застосування методу «ковзного вікна» до його формування;
- вдосконалено процес архівації за рахунок застосування вдосконаленого процесу формування основного словника.

Практичне значення одержаних результатів полягає у такому: розроблено новий метод архівації однотипних файлів, вдосконалено процес формування основного словника за рахунок застосування до його формування методу «ковзного вікна»; розроблено алгоритм роботи формування основного словника однотипних файлів, який дозволяє ефективніше стискати велику кількість однотипних файлів; розроблено програмний засіб для архівації однотипних файлів.

В рамках даної роботи виявлено, що розроблений метод архівації однотипних файлів гарно працює в умовах наявності не менше 10 файлів на вході методу, а при наявності більшої кількості (100 і більше) однотипних файлів значення Arh_m росте. Запропонований метод дає приріст середнього значення процесу архівації на 10,85%, а зі збільшенням кількості однотипних файлів для процесу архівації цей показник може все більше зростати.

Список літератури

[1] Семѐнов Ю. Телекоммуникационные технологи / Семенов Ю. – М.: Бином. Лаборатория знаний, 2007. – 640с.
[2] Luzhetsky, V.A., Savytska, L.A., Troianovska, T.I. Adaptive compression methods of data based on Fibonacci linear forms. G.2017 Proceedings of SPIE - The International Society for Optical Engineering.
[3] Сергеенко В. Сжатие данных, речи, звука и изображений в телекоммуникационных системах / Сергеенко В., Баринов В. – М.: КудицОбраз, 2009. – 360с.
[4] Молдовян А. Криптография / Молдовян А., Советов Б. – М.: Лань, 2000. – 224с.
[5] Макконнелл С. Совершенный код / Макконнелл С. – СПб.: Питер, 2007. – 896с.
[6] Azarov, O.D., Troianovska, T.I., Savytska, L.A., Kozbekova, A., Sagymbekova, A. Quality of content delivery in computer specialists training system. G.2017 Proceedings of SPIE - The International Society for Optical Engineering.

Стаття надійшла: 12.12.19.

References

[1] Semènov Yu. Telekommunykatsyonnыe tekhnolohy / Semenov Yu. – M.: Bynom. Laboratoryia znanyi, 2007. – 640s.
[2] Luzhetsky, V.A., Savytska, L.A., Troianovska, T.I. Adaptive compression methods of data based on Fibonacci linear forms. G.2017 Proceedings of SPIE - The International Society for Optical Engineering.
[3] Serheenko V. Szhatye dannыkh, rechy, zvuka y yzobrazhenyi v telekommunykatsyonnыkh systemakh / Serheenko V., Barynov V. – M.: KudytsObraz, 2009. – 360s.
[4] Moldovian A. Kryptohrafyia / Moldovian A., Sovetov B. – M.: Lan, 2000. – 224s.
[5] Makkonnell S. Sovershennыi kod / Makkonnell S. – SPb.: Pyter, 2007. – 896s.
[6] Azarov, O.D., Troianovska, T.I., Savytska, L.A., Kozbekova, A., Sagymbekova, A. Quality of content delivery in computer specialists training system. G.2017 Proceedings of SPIE - The International Society for Optical Engineering.

Відомості про авторів

Савицька Людмила Анатоліївна, к. т. н., доцент кафедри обчислювальної техніки, ВНТУ, кафедра обчислювальної техніки, Вінниця, Хмельницьке шосе, 95

Коробейнікова Тетяна Іванівна, к.т.н., доцент кафедри обчислювальної техніки, ВНТУ, кафедра обчислювальної техніки, Вінниця, Хмельницьке шосе, 95

Чирва Павло Васильович, магістр кафедри обчислювальної техніки, ВНТУ, кафедра обчислювальної техніки, Вінниця, Хмельницьке шосе, 95

Л. А. Савицька, Т. И. Коробейникова, П. В. Чирва

МЕТОД ТА КРОСПЛАТФОРМЕНИЙ ЗАСІБ АРХІВАЦІЇ ОДНОТИПНИХ ФАЙЛІВ

Вінницький національний технічний університет, м. Вінниця

L.A. Savytska, T.I. Korobeinikova, P. V. Chyrva

METHOD AND CROSS-PLATFORM FILE ARCHIVING TOOL

Vinnytsia National Technical University, Vinnitsa